A Appendix

A.1 Mathematical miscellany

In many cases we would like to bound a summation using an integral.

Lemma 3. *For $x \geq 0$, we have*

$$\sum_{i=a}^{b} i^x \leq \int_{a}^{b+1} i^x di = \frac{(b+1)^{x+1} - a^{x+1}}{x+1} \tag{11}$$

$$\sum_{i=a}^{b} i^x \geq \int_{a-1}^{b} i^x di = \frac{b^{x+1} - (a-1)^{x+1}}{x+1} \tag{12}$$

For $x < 0$ and $x \neq -1$, we have

$$\sum_{i=a}^{b} i^x \leq \int_{a-1}^{b} i^x di = \frac{b^{x+1} - (a-1)^{x+1}}{x+1} \tag{13}$$

$$\sum_{i=a}^{b} i^x \geq \int_{a}^{b+1} i^x di = \frac{(b+1)^{x+1} - a^{x+1}}{x+1} \tag{14}$$

For $x = -1$, we have

$$\sum_{i=a}^{b} i^x \leq \int_{a-1}^{b} i^x di = \log \frac{b}{a-1} \tag{15}$$

$$\sum_{i=a}^{b} i^x \geq \int_{a}^{b+1} i^x di = \log \frac{b+1}{a} \tag{16}$$

The sequence $\{i^x\}$ is increasing when $x > 0$ and is decreasing when $x < 0$. The proof follows directly from applying standard technique of bounding summation with integral.

A.2 Proofs from Section 2

Proof. (Of Theorem 1) Consider an oracle $\mathcal{G}$ implemented based on a dataset D of size T. Given any sequence $w_1, w_2, \ldots, w_T$, the *disguised version* of D output by $\mathcal{G}$ is the sequence of gradients $\mathcal{G}(w_1), \ldots, \mathcal{G}(w_T)$. Suppose that the oracle accesses the data in a (random) order specified by a permutation π; for any t, any $x, x' \in \mathcal{X}$, $y, y' \in \{1, -1\}$, we have

$$\begin{aligned}
\frac{\rho(\mathcal{G}(w_t) = g | (x_{\pi(t)}, y_{\pi(t)}) = (x, y))}{\rho(\mathcal{G}(w_t) = g | (x_{\pi(t)}, y_{\pi(t)}) = (x', y'))} =& \frac{\rho(Z_t = g - \lambda w - \nabla \ell(w, x, y))}{\rho(Z_t = g - \lambda w - \nabla \ell(w, x', y'))} \\
=& \frac{e^{-(\epsilon/2)\|g - \lambda w - \nabla \ell(w,x,y)\|}}{e^{-(\epsilon/2)\|g - \lambda w - \nabla \ell(w,x',y')\|}} \\
\leq& \exp\left((\epsilon/2)(\|\nabla \ell(w, x, y)\| + \|\nabla \ell(w, x', y')\|)\right) \\
\leq& \exp(\epsilon) .
\end{aligned}$$

The first inequality follows from the triangle inequality, and the last step follows from the fact that $\|\nabla \ell(w, x, y)\| \leq 1$. The privacy proof follows.

For the rest of the theorem, we consider a slightly generalized version of SGD that includes mini-batch updates. Suppose the batch size is b; for standard SGD, $b = 1$. For a given t, we call $\mathcal{G}(w_t)$ b successive times to obtain noisy gradient estimates $g_1(w_t), \ldots, g_b(w_t)$; these are gradient estimates at w_t but are based on separate (private) samples. The SGD update rule is:

$$w_{t+1} = \Pi_{\mathcal{W}}\left(w_t - \frac{\eta_t}{b}(g_1(w_t) + \ldots + g_b(w_t))\right).$$

For any i, $\mathbb{E}[g_i(w_t)] = \lambda w + \mathbb{E}[\nabla \ell(w, x, y)]$, where the first expectation is with respect to the data distribution and the noise, and the second is with respect to the data distribution; the unbiasedness result follows.

We now bound the norm of the noisy gradient calculated from a batch. Suppose that the oracle accesses the dataset D in an order π. Then, $g_i(w_t) = \lambda w + \nabla \ell(w_t, x_{\pi((t-1)b+i)}, y_{\pi((t-1)b+i)}) + Z_{(t-1)b+i}$. Expanding on the expression for the expected squared norm of the gradient, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{b}(g_1(w_t) + \ldots + g_b(w_t))\right\|^2\right] =& \mathbb{E}\left[\left\|\lambda w + \frac{1}{b}\sum_{i=1}^{b} \nabla \ell(w_t, x_{\pi((t-1)b+i)}, y_{\pi((t-1)b+i)})\right\|^2\right] \\ &+ \frac{2}{b}\mathbb{E}\left[\left(\lambda w + \frac{1}{b}\sum_{i=1}^{b} \nabla \ell(w_t, x_{\pi((t-1)b+i)}, y_{\pi((t-1)b+i)})\right) \cdot \left(\sum_{i=1}^{b} Z_{(t-1)b+i}\right)\right] \\ &+ \frac{1}{b^2}\mathbb{E}\left[\left\|\sum_{i=1}^{b} Z_{(t-1)b+i}\right\|^2\right] \qquad (17)\end{aligned}$$

We now look at the three terms in (17) separately.
The first term can be further expanded to:

$$\begin{aligned}\mathbb{E}\left[\|\lambda w\|^2\right] + \mathbb{E}\left[\left\|\frac{1}{b^2}\sum_{i=1}^{b} \nabla \ell(w_t, x_{\pi((t-1)b+i)}, y_{\pi((t-1)b+i)})\right\|^2\right] \\ + 2\lambda w \cdot \left(\sum_{i=1}^{b} \mathbb{E}\left[\nabla \ell(w_t, x_{\pi((t-1)b+i)}, y_{\pi((t-1)b+i)})\right]\right) \qquad (18)\end{aligned}$$

The first term in (18) is at most $\lambda^2 \max_{w \in \mathcal{W}} \|w\|^2$, which is at most 1. The second term is at most $\max_w \lambda \|w\| \cdot \max_{w,x,y} \|\nabla \ell(w, x, y)\| \leq 1$, and the third term is at most 2. Thus, the first term in (17) is at most 4. Notice that this upper bound can be pretty loose compare to the average $\left\|\lambda w + \frac{1}{b}\sum_{i=1}^{b} \nabla \ell(w_t, x_{\pi((t-1)b+i)}, y_{\pi((t-1)b+i)})\right\|^2$ values seen in experiment. This leads to a loose estimation of the noise level for oracle $\mathcal{G}^{\text{DP}}$.
To bound the second term in (17), observe that for all i, $Z_{(t-1)b+i}$ is independent of any $Z_{(t-1)b+i'}$ when $i \neq i'$, as well as of the dataset. Combining this with the fact that $\mathbb{E}[Z_\tau] = 0$ for any τ, we get that this term is 0.

To bound the third term in (17), we have:

$$\begin{aligned}
\frac{1}{b^2}\mathbb{E}\left[\left\|\sum_{t\in B} Z_t\right\|_2^2\right] &= \frac{1}{b^2}\mathbb{E}\left[\sum_{t\in B}\|Z_t\|_2^2 + \sum_{t\in B, s\in B, t\neq s} Z_t\cdot Z_s\right] \\
&= \frac{1}{b^2}\sum_{t\in B}\mathbb{E}\left[\|Z_t\|_2^2\right] + \frac{1}{b^2}\sum_{t\in B, s\in B, t\neq s}\mathbb{E}\left[Z_t\right]\cdot\mathbb{E}\left[Z_s\right] \\
&= \frac{1}{b^2}\sum_{t\in B}\mathbb{E}\left[\|Z_t\|_2^2\right],
\end{aligned}$$

where the first equality is from the linearity of expectation and the last two equalities is from the fact that Z_i is independently drawn zeros mean vector. Because Z_t follows $\rho(Z_t = z) \propto e^{-(\epsilon/2)\|z\|}$, we have

$$\rho(\|Z_t\| = x) \propto x^{d-1}e^{-(\epsilon/2)x},$$

which is a Gamma distribution. For $X \sim \mathrm{Gamma}(k,\theta)$, $\mathbb{E}\left[X\right] = k\theta$ and $\mathrm{Var}\left(X\right) = k\theta^2$. Also, by property of expectation, $\mathbb{E}\left[X^2\right] = (\mathbb{E}\left[X\right])^2 + \mathrm{Var}\left(X\right)$. We then have $\mathbb{E}\left[\|Z_t\|_2^2\right] = \dfrac{4(d^2+d)}{\epsilon^2}$ and the whole term equals to $\dfrac{4(d^2+d)}{\epsilon^2 b}$.

Combining the three bounds together, we have a final bound of $4 + \dfrac{4(d^2+d)}{\epsilon^2 b}$. The lemma follows. $\square$

A.3 Proofs from Section 3

Proof. (of Theorem 2) Let the superscripts CF, NF and AO indicate the iterates for the CF, NF and AO algorithms. Let w_1 denote the initial point of the optimization. Let $(x_t^{\mathsf{O}}, y_t^{\mathsf{O}})$ be the data used under order $\mathsf{O} = \mathsf{CF}, \mathsf{NF}$ or AO to update w at time t, Z_i^{O} be the noise added to the exact gradient by $\mathcal{G}_{\mathsf{C}}$ or $\mathcal{G}_{\mathsf{N}}$, depending on which oracle is used by O at t and w_t^{O} be the w obtained under order O at time t. Then by expanding the expression for w_t in terms of the gradients, we have

$$w_{T+1}^{\mathsf{O}} = w_1\prod_{i=1}^{T}(1-\eta_t\lambda) - \sum_{t=1}^{T}\eta_t\left(\prod_{s=t+1}^{T}(1-\eta_s\lambda)\right)(y_t^{\mathsf{O}}x_t^{\mathsf{O}} + Z_t^{\mathsf{O}}). \tag{19}$$

Similarly, if $v_1 = w_1$, we have

$$v_{T+1}^{\mathsf{O}} = w_1\prod_{i=1}^{T}(1-\eta_t\lambda) - \sum_{t=1}^{T}\eta_t\left(\prod_{s=t+1}^{T}(1-\eta_s\lambda)\right)y_t^{\mathsf{O}}x_t^{\mathsf{O}}. \tag{20}$$

Define

$$\Delta_t = \eta_t\prod_{s=t+1}^{\top}(1-\eta_s\lambda).$$

Taking the expected squared difference between (19) from (20), we obtain

$$\begin{aligned}
\mathbb{E}\left[\|v_{T+1}^{\mathsf{O}} - w_{T+1}^{\mathsf{O}}\|^2\right] &= \mathbb{E}\left[\left\|\sum_{t=1}^{T} \eta_t \left(\prod_{s=t+1}^{T} (1-\eta_s\lambda)\right) Z_t^{\mathsf{O}}\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\sum_{t=1}^{T} \Delta_t Z_t^{\mathsf{O}}\right\|^2\right] \\
&= \sum_{t=1}^{T} \Delta_t^2 \mathbb{E}\left[\|Z_t^{\mathsf{O}}\|^2\right], \qquad (21)
\end{aligned}$$

where the second step follows because the Z_i^{O}'s are independent.
If $\eta_t = c/t$, then

$$\Delta_t = \frac{c}{t} \prod_{s=t+1}^{\top} \left(1 - \frac{c\lambda}{s}\right).$$

Therefore

$$\frac{\Delta_{t+1}^2}{\Delta_t^2} = \left(\frac{\frac{c}{t+1} \prod_{s=t+2}^{\top} \left(1-\frac{c\lambda}{s}\right)}{\frac{c}{t} \prod_{s=t+1}^{\top} \left(1-\frac{c\lambda}{s}\right)}\right)^2 = \left(\frac{t}{(t+1)\left(1-\frac{c\lambda}{t+1}\right)}\right)^2 = \left(\frac{1}{1+\frac{1-c\lambda}{t}}\right)^2,$$

which is smaller than 1 if $c < 1/\lambda$, equal to 1 if $c = 1/\lambda$, and greater than 1 if $c > 1/\lambda$. Therefore Δ_t is decreasing if $c < 1/\lambda$ and is increasing if $c > 1/\lambda$.
If Δ_t is decreasing, then (21) is minimized if $\mathbb{E}\left[\|Z_t^{\mathsf{O}}\|^2\right]$ is increasing; if Δ_t is increasing, then (21) is minimized if $\mathbb{E}\left[\|Z_t^{\mathsf{O}}\|^2\right]$ is decreasing; and if Δ_t is constant, then (21) is the same under any order of $\mathbb{E}\left[\|Z_t^{\mathsf{O}}\|^2\right]$.
Therefore for $c < 1/\lambda$,

$$\mathbb{E}\left[\left\|v_{T+1}^{\mathsf{CF}} - w_{T+1}^{\mathsf{CF}}\right\|^2\right] \le \mathbb{E}\left[\left\|v_{T+1}^{\mathsf{AO}} - w_{T+1}^{\mathsf{AO}}\right\|^2\right] \le \mathbb{E}\left[\left\|v_{T+1}^{\mathsf{NF}} - w_{T+1}^{\mathsf{NF}}\right\|^2\right].$$

For $c = 1/\lambda$,

$$\mathbb{E}\left[\left\|v_{T+1}^{\mathsf{CF}} - w_{T+1}^{\mathsf{CF}}\right\|^2\right] = \mathbb{E}\left[\left\|v_{T+1}^{\mathsf{AO}} - w_{T+1}^{\mathsf{AO}}\right\|^2\right] = \mathbb{E}\left[\left\|v_{T+1}^{\mathsf{NF}} - w_{T+1}^{\mathsf{NF}}\right\|^2\right].$$

For $c > 1/\lambda$,

$$\mathbb{E}\left[\left\|v_{T+1}^{\mathsf{CF}} - w_{T+1}^{\mathsf{CF}}\right\|^2\right] \ge \mathbb{E}\left[\left\|v_{T+1}^{\mathsf{AO}} - w_{T+1}^{\mathsf{AO}}\right\|^2\right] \ge \mathbb{E}\left[\left\|v_{T+1}^{\mathsf{NF}} - w_{T+1}^{\mathsf{NF}}\right\|^2\right].$$

□

A.4 Proofs from Section 4

Recall that we have oracles $\mathcal{G}_1, \mathcal{G}_2$ based on data sets D_1 and D_2. The fractions of data in each data set are $\beta_1 = \frac{|D_1|}{|D_1|+|D_2|}$ and $\beta_2 = \frac{|D_2|}{|D_1|+|D_2|}$, respectively.

A.4.1 Proof of Theorem 3

Theorem 3 is a corollary of the following Lemma.

Lemma 4. *Consider the SGD algorithm that follows Algorithm 1. Suppose the objective function is λ-strongly convex, and define $\mathcal{W} = \{w : \|w\| \leq B\}$. If $2\lambda c_1 > 1$ and $i_0 = \lceil 2c_1\lambda \rceil$, then we have the following two cases:*

1. *If $2\lambda c_2 \neq 1$,*

$$\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] \leq \left(4\Gamma_1^2 \frac{\beta_1^{2\lambda c_2 - 1} c_1^2}{2\lambda c_1 - 1} + 4\Gamma_2^2 \frac{c_2^2(1 - \beta_1^{2\lambda c_2 - 1})}{2\lambda c_2 - 1}\right) \cdot \frac{1}{T} + \mathcal{O}\left(\frac{1}{T^{\min(2\lambda c_1, 2)}}\right)$$

2. *If $2\lambda c_2 = 1$,*

$$\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] \leq \left(4\Gamma_1^2 \frac{\beta_1^{2\lambda c_2 - 1} c_1^2}{2\lambda c_1 - 1} + 4\Gamma_2^2 c_2^2 \log \frac{1}{\beta_1}\right) \cdot \frac{1}{T} + \mathcal{O}\left(\frac{1}{T^{\min(2\lambda c_1, 2)}}\right)$$

We first begin with a lemma which follows from arguments very similar to those made in Rakhlin et al. (2012).

Lemma 5. *Let w^* be the optimal solution to $\mathbb{E}[f(w)]$. Then,*

$$\mathbb{E}_{1,\ldots,t}\left[\|w_{t+1} - w^*\|^2\right] \leq (1 - 2\lambda\eta_t)\mathbb{E}_{1,\ldots,t}\left[\|w_t - w^*\|^2\right] + \eta_t^2\gamma_t^2.$$

where the expectation is taken wrt the oracle as well as sampling from the data distribution.

Proof. (Of Lemma 5) By strong convexity of f, we have

$$f(w') \geq f(w) + g(w)^\top(w' - w) + \frac{\lambda}{2}\|w - w'\|^2. \tag{22}$$

Then by taking $w = w_t$, $w' = w^*$ we have

$$g(w_t)^\top(w_t - w^*) \geq f(w_t) - f(w^*) + \frac{\lambda}{2}\|w_t - w^*\|^2. \tag{23}$$

And similarly by taking $w' = w_t$, $w = w^*$, we have

$$f(w_t) - f(w^*) \geq \frac{\lambda}{2}\|w_t - w^*\|^2. \tag{24}$$

By the update rule and convexity of $\mathcal{W}$, we have

$$\begin{aligned}
\mathbb{E}_{1,\ldots,t}\left[\|w_{t+1} - w^*\|^2\right] &= \mathbb{E}_{1,\ldots,t}\left[\|\Pi_{\mathcal{W}}\left(w_t - \eta_t\hat{g}(w_t)\right) - w^*\|^2\right] \\
&\leq \mathbb{E}_{1,\ldots,t}\left[\|w_t - \eta_t\hat{g}(w_t) - w^*\|^2\right] \\
&= \mathbb{E}_{1,\ldots,t}\left[\|w_t - w^*\|^2\right] - 2\eta_t\mathbb{E}_{1,\ldots,t}\left[\hat{g}(w_t)^\top(w_t - w^*)\right] \eta_t^2\mathbb{E}_{1,\ldots,t}\left[\|\hat{g}(w_t)\|^2\right].
\end{aligned}$$

Consider the term $\mathbb{E}_{1,\ldots,t}\left[\hat{g}(w_t)^\top(w_t - w^*)\right]$, where the expectation is taken over the randomness from time 1 to t. Since w_t is a function of the samples used from time 1 to $t-1$, it is independent of the sample used at t. So we have

$$\begin{aligned}\mathbb{E}_{1,\dots,t}\left[\|w_{t+1}-w^*\|^2\right] &\leq \mathbb{E}_{1,\dots,t}\left[\hat{g}(w_t)^\top(w_t-w^*)\right]\\ &= \mathbb{E}_{1,\dots,t-1}\left[\mathbb{E}_t[\hat{g}(w_t)^\top(w_t-w^*)|w_t]\right]\\ &= \mathbb{E}_{1,\dots,t-1}\left[\mathbb{E}_t[\hat{g}(w_t)^\top|w_t](w_t-w^*)\right]\\ &= \mathbb{E}_{1,\dots,t-1}\left[g(w_t)^\top(w_t-w^*)\right].\end{aligned}$$

We have the following upper bound:

$$\begin{aligned}\mathbb{E}_{1,\dots,t}\left[\|w_{t+1}-w^*\|^2\right] \leq \mathbb{E}_{1,\dots,t}\left[\|w_t-w^*\|^2\right] &- 2\eta_t\mathbb{E}_{1,\dots,t-1}\left[g(w_t)^\top(w_t-w^*)\right]\\ &+ \eta_t^2\mathbb{E}_{1,\dots,t}\left[\|\hat{g}(w_t)\|^2\right].\end{aligned}$$

By (23) and the bound $\mathbb{E}\left[\|\hat{g}(w_t)\|^2\right] \leq \gamma_t^2$, we have

$$\mathbb{E}_{1,\dots,t}\left[\|w_{t+1}-w^*\|^2\right] \leq \mathbb{E}_{1,\dots,t}\left[\|w_t-w^*\|^2\right] - 2\eta_t\mathbb{E}_{1,\dots,t-1}\left[f(w_t)-f(w^*)+\frac{\lambda}{2}\|w_t-w^*\|^2\right] + \eta_t^2\gamma_t^2.$$

Then by (24) and the fact that w_t is independent of the sample used in time t, we have the following recursion:

$$\mathbb{E}_{1,\dots,t}\left[\|w_{t+1}-w^*\|^2\right] \leq (1-2\lambda\eta_t)\mathbb{E}_{1,\dots,t}\left[\|w_t-w^*\|^2\right] + \eta_t^2\gamma_t^2.$$

□

Proof. (Of Lemma 4) Let $g(w)$ be the true gradient $\nabla f(w)$ and $\hat{g}(w)$ be the unbiased noisy gradient provided by the oracle $\mathcal{G}_1$ or $\mathcal{G}_2$, whichever is queried. From Lemma 5, we have the following recursion:

$$\mathbb{E}_{1,\dots,t}\left[\|w_{t+1}-w^*\|^2\right] \leq (1-2\lambda\eta_t)\mathbb{E}_{1,\dots,t}\left[\|w_t-w^*\|^2\right] + \eta_t^2\gamma_t^2.$$

Let i_0 be the smallest positive integer such that $2\lambda\eta_{i_0} < 1$, i.e, $i_0 = \lceil 2c_1\lambda\rceil$. Notice that for fixed step size constant c and λ, i_0 would be a fixed constant. Therefore we assume that $i_0 < \beta T$. Using the above inequality inductively, and substituting $\gamma_t = \Gamma_1$ for $t \leq \beta_1 T$ and $\gamma_t = \Gamma_2$ for $t > \beta_1 T$, we have

$$\begin{aligned}\mathbb{E}_{1,\dots,T}\left[\|w_{T+1}-w^*\|^2\right] \leq \prod_{i=i_0}^{\beta_1 T}(1-2\lambda\eta_i)&\prod_{i=\beta_1 T+1}^{T}(1-2\lambda\eta_i)\,\mathbb{E}_{1,\dots,T}\left[\|w_{i_0}-w^*\|^2\right]\\ &+\Gamma_1^2\prod_{i=\beta_1 T+1}^{T}(1-2\lambda\eta_i)\sum_{i=i_0}^{\beta_1 T}\eta_i^2\prod_{j=i+1}^{\beta_1 T}(1-2\lambda\eta_j)\\ &+\Gamma_2^2\sum_{i=\beta_1 T+1}^{T}\eta_i^2\prod_{j=i+1}^{T}(1-2\lambda\eta_j).\end{aligned}$$

By substituting $\eta_t = \dfrac{c_1}{t}$ for D_1 and $\eta_t = \dfrac{c_2}{t}$ for D_2, we have

$$\begin{aligned}
\mathbb{E}_{1,\ldots,T}\left[\|w_{T+1} - w^*\|^2\right] \leq \prod_{i=i_0}^{\beta_1 T}\left(1 - \frac{2\lambda c_1}{i}\right) &\prod_{i=\beta_1 T+1}^{T}\left(1 - \frac{2\lambda c_2}{i}\right)\mathbb{E}_{1,\ldots,T}\left[\|w_{i_0} - w^*\|^2\right] \\
&+ \Gamma_1^2 \prod_{i=\beta_1 T+1}^{T}\left(1 - \frac{2\lambda c_2}{i}\right)\sum_{i=i_0}^{\beta_1 T}\frac{c_1^2}{i^2}\prod_{j=i+1}^{\beta_1 T}\left(1 - \frac{2\lambda c_1}{j}\right) \\
&+ \Gamma_2^2 \sum_{i=\beta_1 T+1}^{T}\frac{c_2^2}{i^2}\prod_{j=i+1}^{T}\left(1 - \frac{2\lambda c_2}{j}\right).
\end{aligned}$$

Applying the inequality $1 - x \leq e^{-x}$ to each of the terms in the products, and simplifying, we get:

$$\begin{aligned}
\mathbb{E}_{1,\ldots,T}\left[\|w_{T+1} - w^*\|^2\right] \leq e^{-2\lambda c_1 \sum_{i=i_0}^{\beta_1 T}\frac{1}{i}} &e^{-2\lambda c_2 \sum_{i=\beta_1 T+1}^{\top}\frac{1}{i}}\mathbb{E}_{1,\ldots,T}\left[\|w_{i_0} - w^*\|^2\right] \\
&+ \Gamma_1^2 e^{-2\lambda c_2 \sum_{i=\beta_1 T+1}^{\top}\frac{1}{i}}\sum_{i=i_0}^{\beta_1 T}\frac{c_1^2}{i^2}e^{-2\lambda c_1 \sum_{j=i+1}^{\beta_1 T}\frac{1}{j}} \\
&+ \Gamma_2^2 \sum_{i=\beta_1 T+1}^{\top}\frac{c_2^2}{i^2}e^{-2\lambda c_2 \sum_{j=i+1}^{\top}\frac{1}{j}}.
\end{aligned} \tag{25}$$

We would like to bound (25) term by term.
A bound we will use later is:

$$e^{2\lambda c_2/\beta_1 T} = 1 + \frac{2\lambda c_2}{\beta_1 T}e^{2\lambda c_2/\beta_1 T'} \leq 1 + \frac{2\lambda c_2}{\beta_1 T}e^{2\lambda c_2/\beta_1}, \tag{26}$$

where the equality is obtained using Taylor's theorem, and the inequality follows because T' is in the range $[1, \infty)$. Now we can bound the three terms in (25) separately.

The first term in (25): We bound this as follows:

$$\begin{aligned}
&e^{-2\lambda c_1 \sum_{i=i_0}^{\beta_1 T}\frac{1}{i}}e^{-2\lambda c_2 \sum_{i=\beta_1 T+1}^{\top}\frac{1}{i}}\mathbb{E}_{1,\ldots,T}\left[\|w_{i_0} - w^*\|^2\right] \\
&\leq e^{-2\lambda c_1 \log\frac{\beta_1 T}{i_0}}e^{-2\lambda c_2 (\log\frac{1}{\beta_1} - \frac{1}{\beta_1 T})}\mathbb{E}_{1,\ldots,T}\left[\|w_{i_0} - w^*\|^2\right] \\
&\leq \left(\frac{i_0}{T}\right)^{2\lambda c_1}\beta_1^{2\lambda(c_2 - c_1)}e^{2\lambda c_2/\beta_1 T}(4B^2) \\
&\leq \left(\frac{i_0}{T}\right)^{2\lambda c_1}\beta_1^{2\lambda(c_2 - c_1)}\left(1 + \frac{2\lambda c_2}{\beta_1 T}e^{2\lambda c_2/\beta_1}\right)4B^2 \\
&= 4B^2 {i_0}^{2\lambda c_1}\beta_1^{2\lambda(c_2 - c_1)}\frac{1}{T^{2\lambda c_1}} + \mathcal{O}\left(\frac{1}{T^{2\lambda c_1 + 1}}\right),
\end{aligned}$$

where the first equality follows from (14).The second inequality follows from $\|w\| \leq B$, $\|w - w'\| \leq \|w\| + \|w'\| \leq 2B$, and bounding expectation using maximum. The third follows from (26).

The second term in (25): We bound this as follows:

$$
\begin{aligned}
\Gamma_1^2 e^{-2\lambda c_2 \sum_{i=\beta_1 T+1}^{\top} \frac{1}{i}} \sum_{i=i_0}^{\beta_1 T} \frac{c_1^2}{i^2} e^{-2\lambda c_1 \sum_{j=i+1}^{\beta_1 T} \frac{1}{j}} &\le \Gamma_1^2 e^{-2\lambda c_2 (\log \frac{1}{\beta_1} - \frac{1}{\beta_1 T})} \sum_{i=i_0}^{\beta_1 T} \frac{c_1^2}{i^2} e^{-2\lambda c_1 \log \frac{\beta_1 T}{i+1}} \\
&= \Gamma_1^2 \beta_1^{2\lambda c_2} e^{2\lambda c_2/\beta_1 T} \sum_{i=i_0}^{\beta_1 T} \frac{c_1^2}{i^2} \left(\frac{i+1}{\beta_1 T} \right)^{2\lambda c_1} \\
&= \Gamma_1^2 \beta_1^{2\lambda(c_2 - c_1)} e^{2\lambda c_2/\beta_1 T} c_1^2 T^{-2\lambda c_1} \sum_{i=i_0}^{\beta_1 T} \frac{(i+1)^{2\lambda c_1}}{i^2} \\
&\le \Gamma_1^2 \beta_1^{2\lambda(c_2 - c_1)} e^{2\lambda c_2/\beta_1 T} c_1^2 T^{-2\lambda c_1} \sum_{i=i_0}^{\beta_1 T} 4(i+1)^{2\lambda c_1 - 2} \\
&\le 4\Gamma_1^2 \beta_1^{2\lambda(c_2 - c_1)} \left(1 + \frac{2\lambda c_2}{\beta_1 T} e^{2\lambda c_2/\beta_1} \right) c_1^2 T^{-2\lambda c_1} \sum_{i=i_0+1}^{\beta_1 T+1} i^{2\lambda c_1 - 2},
\end{aligned}
\tag{27}
$$

where the first inequality follows from (14), the second inequality follows from $(1+\frac{1}{i})^2 \le (1+\frac{1}{1})^2 = 4$, and the last inequality follows from (26).

Bounding summation using integral following (13) and (11) of Lemma 3, if $2\lambda c_1 > 1$, the term on the right hand side would be in the order of $\mathcal{O}(1/T)$; if $2\lambda c_1 = 1$, it would be $\mathcal{O}(\log T/T)$; if $2\lambda c_1 < 1$, it would be $\mathcal{O}(1/T^{2\lambda c_1})$. Therefore to minimize the bound in terms of order, we would choose c_1 such that $2\lambda c_1 > 1$. To get an upper bound of the summation in (27), using (13) of Lemma 3, for $2\lambda c_1 < 2$,

$$
\sum_{j=i_0+1}^{\beta_1 T+1} i^{2\lambda c_1 - 2} = \sum_{j=i_0+1}^{\beta_1 T} i^{2\lambda c_1 - 2} + (\beta_1 T + 1)^{2\lambda c_1 - 2} \le \frac{(\beta_1 T)^{2\lambda c_1 - 1}}{2\lambda c_1 - 1} + \mathcal{O}(T^{2\lambda c_1 - 2}).
$$

For $2\lambda c_1 > 2$, using (11) of Lemma 3,

$$
\sum_{j=i_0+1}^{\beta_1 T+1} i^{2\lambda c_1 - 2} = \sum_{j=i_0+1}^{\beta_1 T-1} i^{2\lambda c_1 - 2} + (\beta_1 T)^{2\lambda c_1 - 2} + (\beta_1 T + 1)^{2\lambda c_1 - 2} \le \frac{(\beta_1 T)^{2\lambda c_1 - 1}}{2\lambda c_1 - 1} + \mathcal{O}(T^{2\lambda c_1 - 2}).
$$

Finally, for $2\lambda c_1 = 2$,

$$
\sum_{j=i_0+1}^{\beta_1 T+1} i^{2\lambda c_1 - 2} = (\beta_1 T + 1) - (i_0 + 1) + 1 = \beta_1 T + \mathcal{O}(1).
$$

Combining the three cases together, we have

$$
\sum_{j=i_0+1}^{\beta_1 T+1} i^{2\lambda c_1 - 2} \le \frac{(\beta_1 T)^{2\lambda c_1 - 1}}{2\lambda c_1 - 1} + \mathcal{O}\left(T^{2\lambda c_1 - 2} \right).
$$

This allows us to further upper bound (27):

$$
\begin{aligned}
&4\Gamma_1^2\beta_1^{2\lambda(c_2-c_1)}\left(1+\frac{2\lambda c_2}{\beta_1 T}e^{2\lambda c_2/\beta_1}\right)c_1^2T^{-2\lambda c_1}\sum_{i=i_0+1}^{\beta_1 T+1} i^{2\lambda c_1-2}\\
&\quad\le 4\Gamma_1^2\beta_1^{2\lambda(c_2-c_1)}\left(1+\frac{2\lambda c_2}{\beta_1 T}e^{2\lambda c_2/\beta_1}\right)c_1^2T^{-2\lambda c_1}\left(\frac{(\beta_1 T)^{2\lambda c_1-1}}{2\lambda c_1-1}+\mathcal{O}\left(T^{2\lambda c_1-2}\right)\right)\\
&\quad=\frac{4\Gamma_1^2c_1^2\beta_1^{2\lambda c_2-1}}{2\lambda c_1-1}\cdot\frac{1}{T}+\mathcal{O}\left(\frac{1}{T^2}\right)+\mathcal{O}\left(\frac{1}{T^3}\right).
\end{aligned}
$$

The last term in (25): We bound this as follows:

$$
\begin{aligned}
&\Gamma_2^2\sum_{i=\beta_1 T+1}^{\top}\frac{c_2^2}{i^2}e^{-2\lambda c_2\sum_{j=i+1}^{\top}\frac{1}{j}}\le\Gamma_2^2\sum_{i=\beta_1 T+1}^{\top}\frac{c_2^2}{i^2}e^{-2\lambda c_2\log\frac{T}{i+1}}\\
&=\Gamma_2^2c_2^2T^{-2\lambda c_2}\sum_{i=\beta_1 T+1}^{\top}\frac{(i+1)^{2\lambda c_2}}{i^2}\le 4\Gamma_2^2c_2^2T^{-2\lambda c_2}\sum_{i=\beta_1 T+1}^{\top}\frac{(i+1)^{2\lambda c_2}}{(i+1)^2}\\
&=4\Gamma_2^2c_2^2T^{-2\lambda c_2}\sum_{i=\beta_1 T+2}^{\top+1}i^{2\lambda c_2-2},
\end{aligned}
\tag{28}
$$

where the first inequality follows from (14) and the last inequality from $(1+\frac{1}{i})^2\le 4$.
If $2\lambda c_2\neq 1$ and $2\lambda c_2\le 2$, using (13) from Lemma 3,

$$
\sum_{j=\beta_1 T+2}^{T+1} i^{2\lambda c_2-2}\le\frac{1-\beta_1^{2\lambda c_2-1}}{2\lambda c_2-1}T^{2\lambda c_2-1}.
$$

If $2\lambda c_2>2$, using (11) from Lemma 3,

$$
\begin{aligned}
\sum_{j=\beta_1 T+2}^{T+1} i^{2\lambda c_2-2}&=\sum_{j=\beta_1 T}^{T-1} i^{2\lambda c_2-2}+T^{2\lambda c_2-2}+(T+1)^{2\lambda c_2-2}-(\beta_1 T+1)^{2\lambda c_2-2}-(\beta_1 T)^{2\lambda c_2-2}\\
&=\frac{1-\beta_1^{2\lambda c_2-1}}{2\lambda c_2-1}T^{2\lambda c_2-1}+\mathcal{O}\left(T^{2\lambda c_2-2}\right).
\end{aligned}
$$

If $2\lambda c_2=2$,

$$
\sum_{j=\beta_1 T+2}^{T+1} i^{2\lambda c_2-2}=\sum_{j=\beta_1 T+2}^{T+1}1=(1-\beta_1)T.
$$

In all three cases we have

$$
\sum_{j=\beta_1 T+2}^{T+1} i^{2\lambda c_2-2}\le\frac{1-\beta_1^{2\lambda c_2-1}}{2\lambda c_2-1}T^{2\lambda c_2-1}+\mathcal{O}\left(T^{2\lambda c_2-2}\right).
$$

Then (28) can be further upper bounded for $2\lambda c_2\neq 1$

$$
4\Gamma_2^2c_2^2T^{-2\lambda c_2}\sum_{i=\beta_1 T+2}^{\top+1}i^{2\lambda c_2-2}\le 4\Gamma_2^2\frac{c_2^2(1-\beta_1^{2\lambda c_2-1})}{2\lambda c_2-1}\cdot\frac{1}{T}+\mathcal{O}\left(\frac{1}{T^2}\right).
\tag{29}
$$

If $2\lambda c_2 = 1$, we have

$$\sum_{j=\beta_1 T+2}^{T+1} i^{2\lambda c_2 - 2} = \sum_{j=\beta_1 T+1}^{T} i^{-1} - (\beta_1 T+1)^{-1} + (T+1)^{-1} \leq \log\frac{1}{\beta_1},$$

and then

$$4\Gamma_2^2 c_2^2 T^{-2\lambda c_2} \sum_{i=\beta_1 T+2}^{\mathrm{T}+1} i^{2\lambda c_2-2} \leq 4\Gamma_2^2 c_2^2 \log\frac{1}{\beta_1}\cdot\frac{1}{T}.$$

which is basically taking the limit as $2\lambda c_2 \to 1$ of the highest order term of (29).

Therefore the summation of the three terms is of order $\mathcal{O}(\frac{1}{T})$ (from the second and third terms), and the constant in the front of the highest order term takes on one of two values:

1. If $2\lambda c_2 \neq 1$,
$$4\Gamma_1^2\frac{c_1^2\beta_1^{2\lambda c_2-1}}{2\lambda c_1 - 1} + 4\Gamma_2^2\frac{c_2^2(1-\beta_1^{2\lambda c_2 - 1})}{2\lambda c_2 - 1}.$$

2. If $2\lambda c_2 = 1$,
$$4\Gamma_1^2\frac{c_1^2\beta_1^{2\lambda c_2-1}}{2\lambda c_1 - 1} + 4\Gamma_2^2 c_2^2\log\frac{1}{\beta_1}.$$

□

A.4.2 Proof of Lemma 1

Proof. (Of Lemma 1) Omitting the constant terms and setting $k_1 = 2\lambda c_1, k_2 = 2\lambda c_2$, we can re-write (10) as $1/T$ times

$$Q(k_1,k_2) = \Gamma_1^2\frac{\beta_1^{k_2-1}k_1^2}{k_1-1} + \Gamma_2^2\frac{(1-\beta_1^{k_2-1})k_2^2}{k_2-1}, \tag{30}$$

with $k_1^* = 2\lambda c_1^* = 2$.
Observe that in this case, $k_2^* \geq 2$. Let $x = k_2 - 1$; then $x \geq 1$. Plugging in $k_1^* = 2$, we can re-write (30) as

$$Q(x) = 4\Gamma_1^2\beta_1^x + \Gamma_2^2(1-\beta_1^x)\left(x + \frac{1}{x} + 2\right). \tag{31}$$

Taking the derivative, we see that

$$Q'(x) = -4\Gamma_1^2\beta_1^x\log(1/\beta_1) + \Gamma_2^2(1-\beta_1^x)\left(1-\frac{1}{x^2}\right) + \Gamma_2^2\left(x+\frac{1}{x}+2\right)\beta_1^x\log(1/\beta_1). \tag{32}$$

Suppose

$$l = \frac{2\log(\Gamma_1/\Gamma_2) + \log\log(1/\beta_1)}{\log(1/\beta_1)}.$$

Observe that $\beta_1^l\log(1/\beta_1) = \frac{\Gamma_2^2}{\Gamma_1^2}$. Plugging $x = l$ in to (32), the first term is $-4\Gamma_2^2$, the second term is at most Γ_2^2, and the third term is at most $\frac{\Gamma_2^4}{\Gamma_1^2}(l + \frac{1}{l} + 2)$. Observe that for any fixed β_1, for large enough Γ_1/Γ_2, $l \geq 1$. Thus, the right hand side of (32) is at most: $-4\Gamma_2^2 + \Gamma_2^2 + \frac{\Gamma_2^4}{\Gamma_1^2}(l+3)$. For fixed β_1, l grows logarithmically in Γ_1/Γ_2, and hence, for large enough Γ_1/Γ_2, $\frac{\Gamma_2^2(l+3)}{\Gamma_1^2}$ will become

arbitrarily small. Therefore, for large enough Γ_1/Γ_2, $Q'(l) < 0$.
Suppose

$$u = \frac{2\log(4\Gamma_1/\Gamma_2) + \log\log(1/\beta_1)}{\log(1/\beta_1)}.$$

Observe that $\beta_1^u \log(1/\beta_1) = \frac{\Gamma_2^2}{16\Gamma_1^2}$. Plugging in $x = u$ to (32), the first term reduces to $-\frac{1}{4}\Gamma_2^2$, the second term is $\Gamma_2^2(1-\beta_1^u)(1-\frac{1}{u^2})$, and the third term is ≥ 0. Observe that as $\Gamma_1/\Gamma_2 \to \infty$ with β_1 fixed, $\beta_1^u \to 0$ and $1/u^2 \to 0$. Thus, for large enough Γ_1/Γ_2, $\Gamma_2^2(1-\beta_1^u)(1-\frac{1}{u^2}) \to \Gamma_2^2$, and therefore $Q'(u) > 0$. Thus, $Q'(x) = 0$ somewhere between l and u and the first part of the lemma follows.
Consider

$$x = \frac{2\log(m\Gamma_1/\Gamma_2) + \log\log(1/\beta_1)}{\log(1/\beta_1)}$$

with $1 \leq m \leq 4$. The first term of (31) is always positive. As for the second term, $x + \frac{1}{x} + 2 \geq x$ for positive x and $\beta_1^x = \frac{\Gamma_2^2}{m^2\Gamma_1^2}\frac{1}{\log(1/\beta_1)}$ is small when Γ_1/Γ_2 is sufficiently large. Therefore for sufficiently large Γ_1/Γ_2, we have $\Gamma_2^2(1-\beta_1^x)(x+\frac{1}{x}+2) \geq \frac{\Gamma_2^2}{2}x$, and thus $Q(x) \geq \frac{\Gamma_2^2}{2}x$, which gives the lower bound. And plugging in $x = l$ gives the upper bound.

□

A.4.3 Proof of Lemma 2

Proof. (Of Lemma 2) Let $k_2 = \epsilon$; then $\epsilon \geq 0$. Plugging in $k_1^* = 2$, we can re-write (30) as

$$Q(\epsilon) = 4\Gamma_1^2\beta_1^{\epsilon-1} + \Gamma_2^2(1-\beta_1^{\epsilon-1})\left(-1+\epsilon+\frac{1}{-1+\epsilon}+2\right). \tag{33}$$

Taking the derivative, we obtain the following:

$$\begin{aligned}
Q'(\epsilon) &= -\,4\Gamma_1^2\beta_1^{\epsilon-1}\log(1/\beta_1) + \Gamma_2^2(1-\beta_1^{\epsilon-1})(1-\frac{1}{(1-\epsilon)^2}) - \frac{\Gamma_2^2\epsilon^2}{1-\epsilon}\beta_1^{\epsilon-1}\log(1/\beta_1) \\
&= -\,\beta_1^{\epsilon-1}\log(1/\beta_1)\left(4\Gamma_1^2 + \frac{\Gamma_2^2\epsilon^2}{1-\epsilon}\right) + \Gamma_2^2(\beta_1^{\epsilon-1}-1)\left(\frac{1}{(1-\epsilon)^2}-1\right) \\
&= -\,\beta_1^{\epsilon-1}\log(1/\beta_1)\left(4\Gamma_1^2 + \frac{\Gamma_2^2\epsilon^2}{1-\epsilon}\right) + \Gamma_2^2(\beta_1^{\epsilon-1}-1)\frac{\epsilon(2-\epsilon)}{(1-\epsilon)^2}.
\end{aligned} \tag{34}$$

For $\epsilon = \frac{\Gamma_1^2}{\Gamma_2^2} \leq 1$, using $1-\beta_1^{1-\epsilon} \leq (1-\epsilon)\log(1/\beta_1)$ and $\beta_1^{\epsilon-1} - 1 = (1-\beta_1^{1-\epsilon})\beta^{\epsilon-1}$, this is at most:

$$-\beta_1^{\epsilon-1}\log(1/\beta_1)\left(4\Gamma_1^2 + \frac{\Gamma_2^2\epsilon^2}{1-\epsilon} - \frac{\Gamma_2^2\epsilon(2-\epsilon)}{1-\epsilon}\right) = -2\Gamma_1^2\beta_1^{\epsilon-1}\log(1/\beta_1).$$

Thus, at $l = \frac{\Gamma_1^2}{\Gamma_2^2}$, $Q'(l) < 0$.
Moreover, for $\epsilon \in [0, \frac{1}{2}]$, $1-\beta_1^{1-\epsilon} \geq \beta_1(1-\epsilon)\log(1/\beta_1)$. Therefore, $Q'(\epsilon)$ is at least:

$$\begin{aligned}
Q'(\epsilon) &\geq -\beta_1^{\epsilon-1}\log(1/\beta_1)\left(4\Gamma_1^2 + \frac{\Gamma_2^2\epsilon^2}{1-\epsilon}\right) + \Gamma_2^2\beta_1^{\epsilon}\log(1/\beta_1)\frac{\epsilon(2-\epsilon)}{1-\epsilon} \\
&\geq \beta_1^{\epsilon-1}\log(1/\beta_1)\left(\frac{\Gamma_2^2\beta\epsilon(2-\epsilon)}{1-\epsilon} - 4\Gamma_1^2 - \frac{\Gamma_2^2\epsilon^2}{1-\epsilon}\right).
\end{aligned}$$

Let $u = \frac{8\Gamma_1^2}{\beta_1 \Gamma_2^2}$; suppose that Γ_2/Γ_1 is large enough such that $u \leq \beta_1/4$. Then, $u(2-u)\beta_1 - u^2 \geq \frac{15u\beta_1}{16}$, and

$$\frac{\Gamma_2^2(u(2-u)\beta_1 - u^2)}{1-u} \geq \frac{15\Gamma_2^2 u\beta_1}{16(1-\beta_1)} \geq \frac{15\Gamma_1^2}{2(1-\beta_1)} \geq 5\Gamma_1^2.$$

Therefore, $Q'(u) > 0$, and thus $Q(\epsilon)$ is minimized at some $\epsilon \in [l, u]$.
For the second part of the lemma, the upper bound is obtained by plugging in $\epsilon = \frac{\Gamma_1}{\Gamma_2}$. For the lower bound, observe that for any $\epsilon \in [l, u]$, $Q(\epsilon) \geq 4\Gamma_1^2 \beta_1^{u-1} \geq 4\Gamma_1^2 \beta_1^{\Gamma_2^2/\beta\Gamma_1^2 - 1}$. $\square$